

Disseny i implementació del Data Warehouse d'una cadena de botigues de roba.

Treball Final de Carrera Bases de Dades Relacionals

Autor: Miguel Ángel Vives Beltrán
Estudiant de Enginyeria Tècnica en Informàtica de Sistemes (ETIS)
Universitat Oberta de Catalunya (UOC)
mvivesb@uoc.edu
Consultor: Manel Rella Ruiz

Índex.

- **Punt de partida.**
- **Requisits funcionals.**
- **Enfocament i mètode seguit.**
- **Planificació.**
- **Arquitectura lògica.**
- **Arquitectura física.**
- **Data Warehouse implantat.**
- **Mòdul estadístic.**
- **Càlcul del mòdul estadístic.**
- **Aspectes de valor afegit implementats.**
- **Càrrega de dades múltiple.**
- **Propostes de millora.**

Punt de partida.

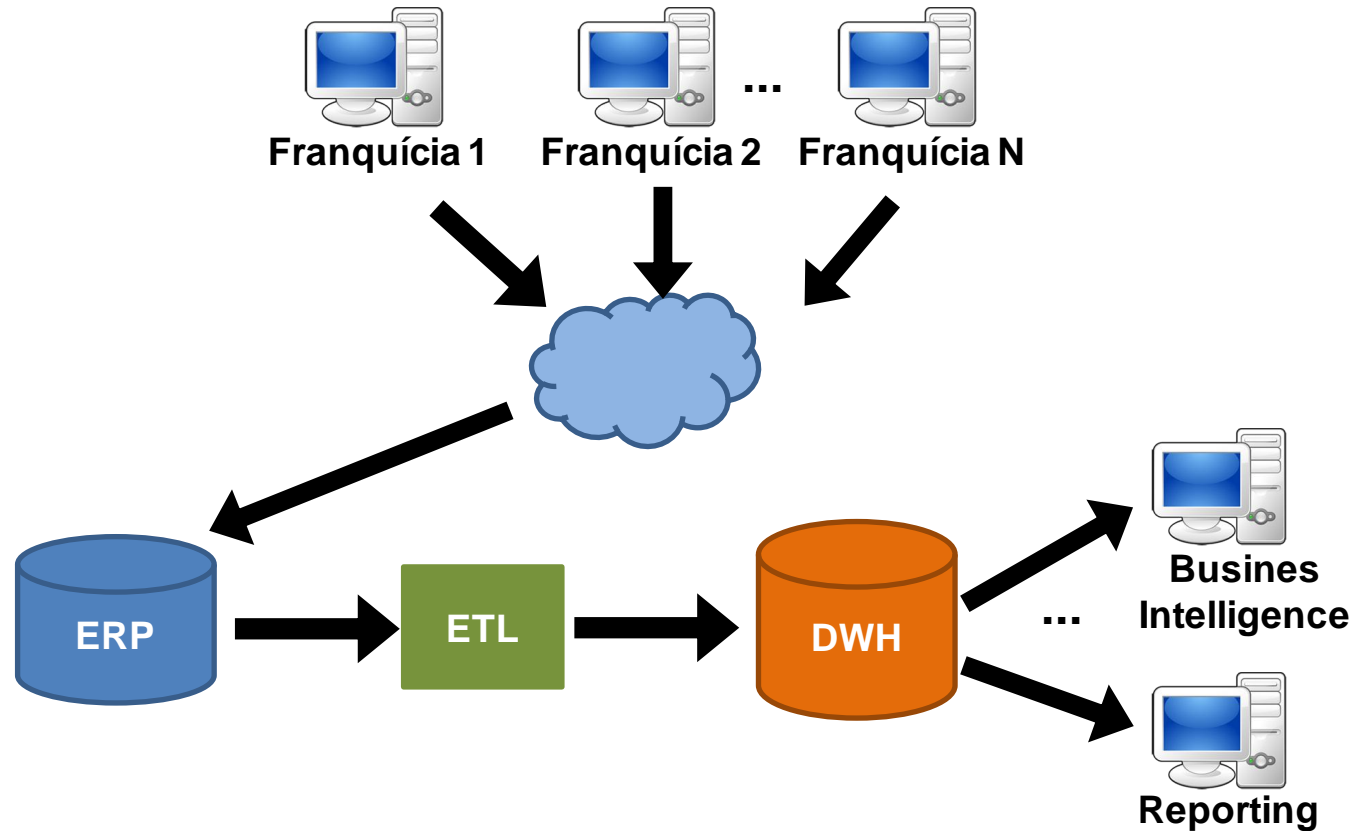
En aquest Treball Final de Carrera de Bases de Dades Relacionals, en endavant TFCBD, s'ha de dissenyar un Data Warehouse, en endavant DWH, per donar cobertura a una sèrie de requisits plantejats.

En el desenvolupament del TFCBD s'han tingut en compte les següents consideracions:

- **El TFCBD es centra únicament en el disseny de la base de dades del DWH:**
 - Disseny i implementació de taules.
 - Disseny i implementació de processos de càrrega.
 - Disseny i implementació del mòdul estadístic.
- **S'han de complir els requisits funcionals plantejats:** Els requisits funcionals es descriuen en la memòria.
- **L'origen de dades del Data Warehouse és el REP de la cadena.**

Punt de partida.

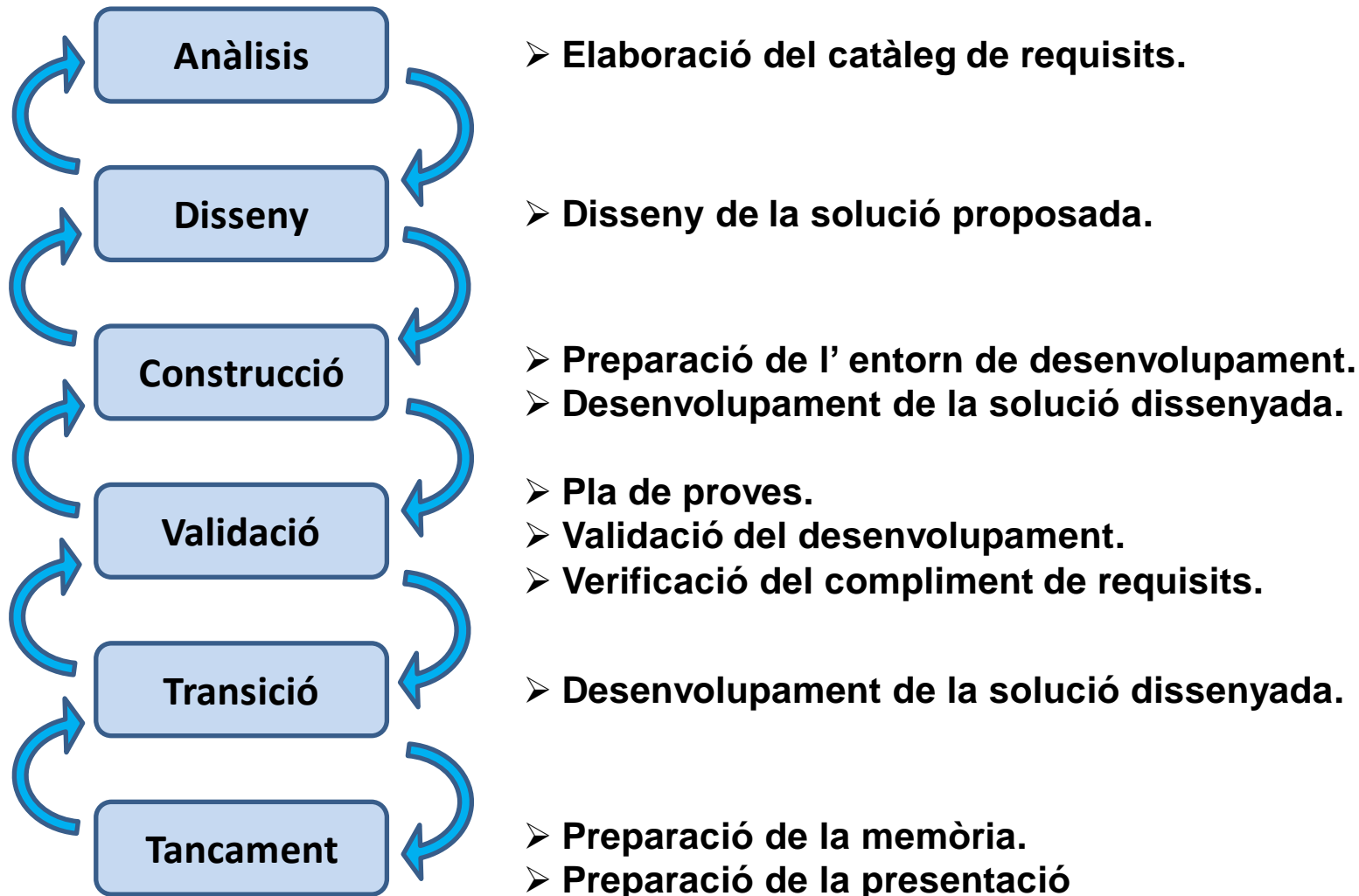
El client serà l'encarregat de integrar el Data Warehouse en la seva infraestructura. A continuació es mostra un exemple d'integració del DWH en la infraestructura de la cadena de botigues de roba:



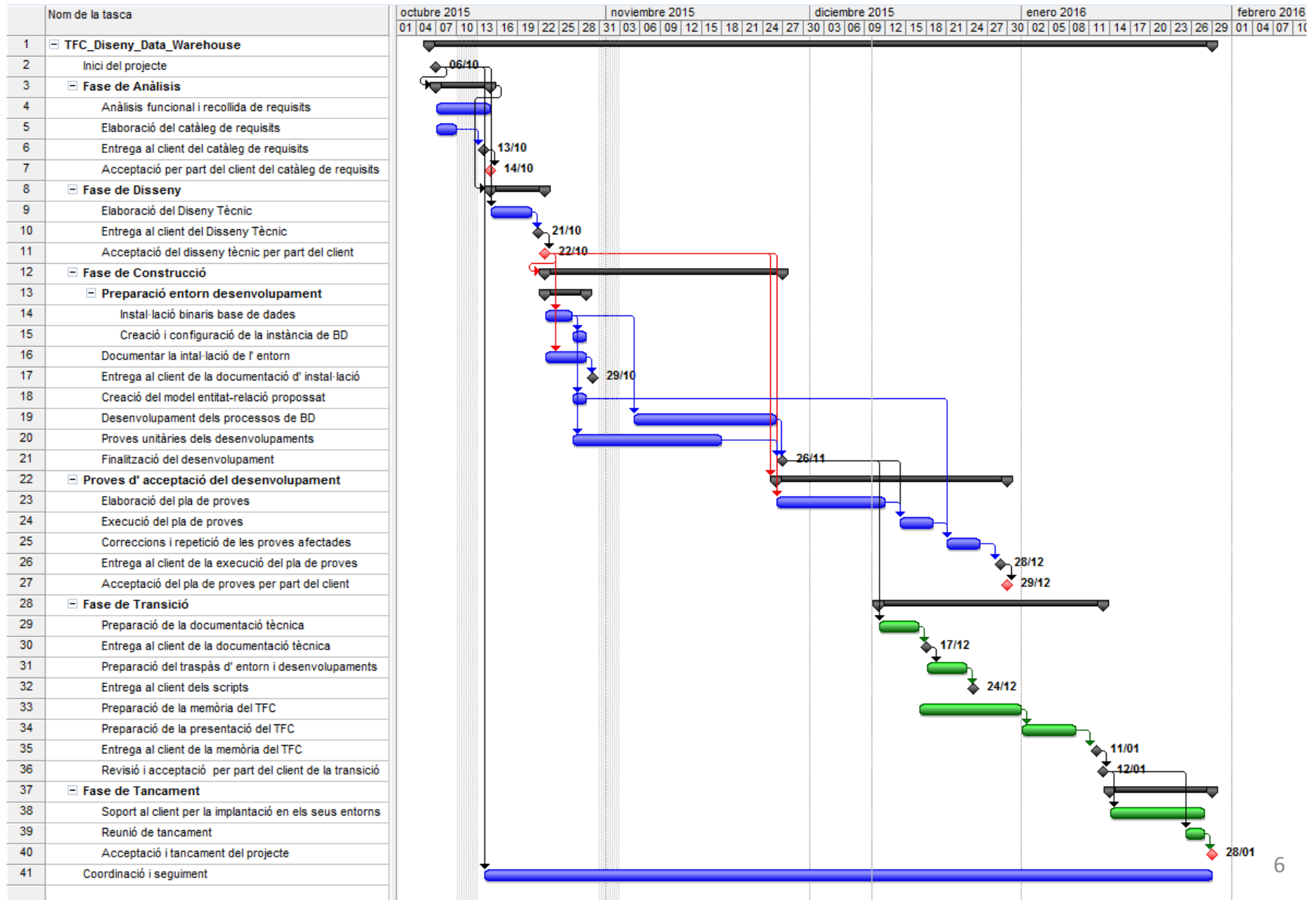
En les properes diapositives es resumirà el treball realitzat per el disseny i implementació del DWH.

Enfocament i mètode seguit.

Per la tipologia del projecte, s'ha seguit una metodologia basada en el model de cascada.

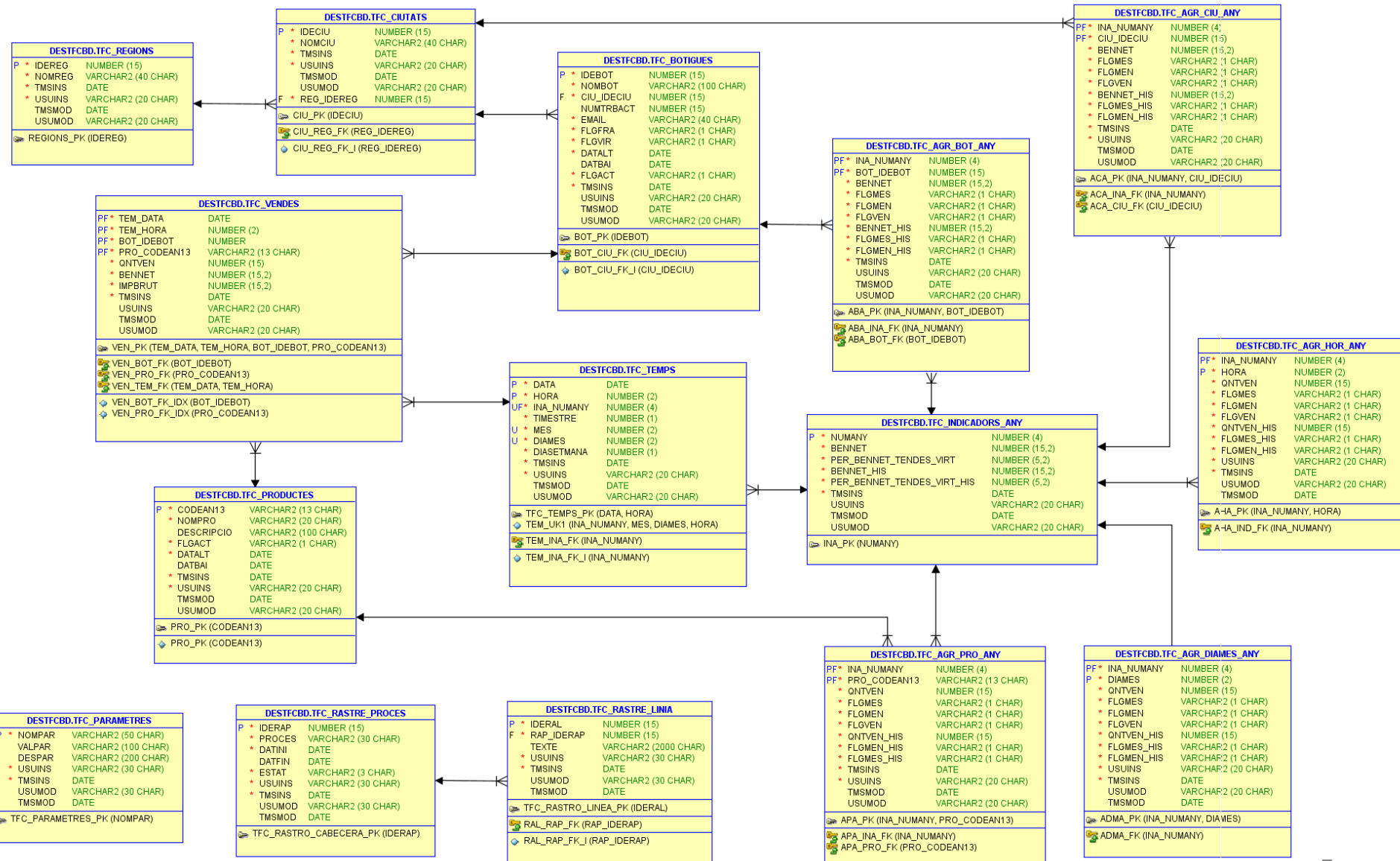


Planificació del projecte.



Arquitectura lògica.

Model entitat - relació



Arquitectura lògica.

Model del Data Warehouse

El Data Warehouse s' ha dissenyat i implementat amb un model de floc de neu amb la taula de fets "TFC_VENDES" com element central i les dimensions normalitzades.

- Taula de fets: **TFC_VENDES**
- Dimensions primàries:
 - Botigues: **TFC_BOTIGUES**
 - Productes: **TFC_PRODUCTES**
 - Temps: **TFC_TEMPS**
- Dimensions secundàries:
 - Regions: **TFC_REGIONS**
 - Ciutats: **TFC_CIUTATS**

A més s' identifiquen les següents agrupacions de taules:

- Mòdul estadístic:
 - **TFC_AGR_CIU_ANY**
 - **TFC_AGR_BOT_ANY**
 - **TFC_AGR_DIAMES_ANY**
 - **TFC_AGR_INDICADORS_ANY**
 - **TFC_AGR_PRO_ANY**
- Taules de sistema:
 - **TFC_PARAMETRES**
 - **TFC_RASTRE_PROCES**
 - **TFC_RASTRE_LINIA**

Arquitectura física.

S' ha treballat en dos entorns de desenvolupament:

- **Base de dades en local:** L' entorn té les següents característiques:
 - **CPU:** Intel(R) Core(TM) i3 CPU M370 2.40 GHz
 - **RAM:** 6 GB
 - **Plataforma:** Windows 7 Professional 64 bits.

S' ha treballat amb el següent programari:

- **Oracle Database Express Edition 11g Release 2**
- **PL/SQL Developer**
- **SQL Developer Data Modeler**

En l' apartat d' annexos de la memòria (Annex 4. Instal·lació Base de Dades), s' adjunta un document amb les passes seguides per dur a terme la instal·lació.

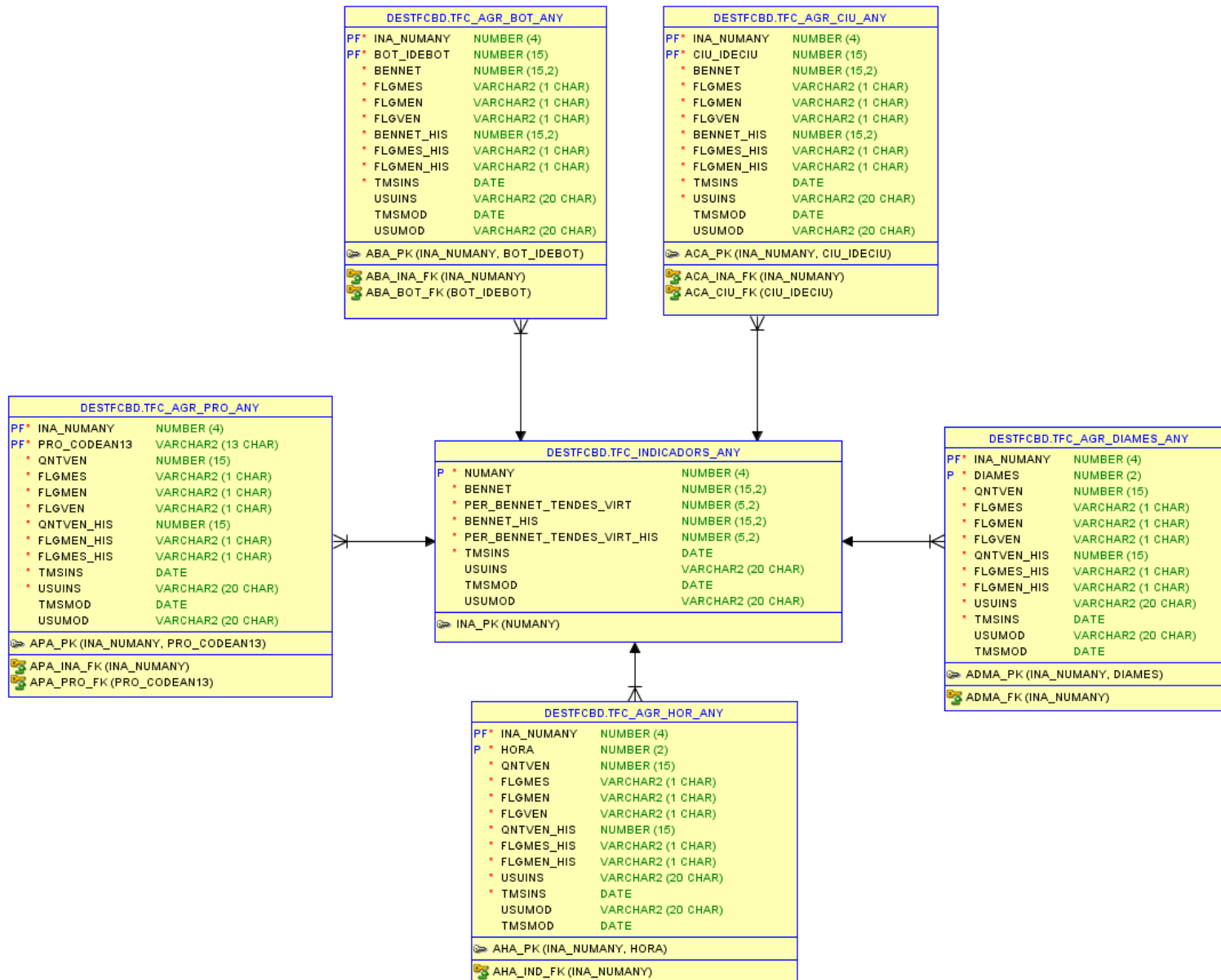
- **Base de dades en el Cloud d' Oracle:** Tal i com es descriu en la memòria (punt "9.8 Annex 8. Aplicació Apex"), s' ha desenvolupat una aplicació en Apex com a exemple d' explotació de les dades. Aquesta aplicació, així com la base de dades que utilitza, estan en el Cloud d' Oracle.

Data Warehouse implantat.

- **Taules:** Es creen les taules representades en l'entitat relació. També es creen taules externes (TFC_LDR*) per el suport a la càrrega múltiple de dades. Es descriuen a la memòria (punt "9.3 Annex 3. Documentació tècnica de taules").
- **Vistes:** Es creen varies vistes per donar cobertura al requeriment R4. Es descriuen en el "punt "9.4 Annex 4. Documentació tècnica de vistes" de la memòria.
- **Paquets PL/SQL:** Punt "9.4 Annex 4. Documentació tècnica de paquets")
 - Gestió del sistema: TFC_UTL, TFC_RAP.
 - Gestió de les dimensions: TFC_BOT, TFC_CIU, TFC_PRO, TFC_REG, TFC_TEM.
 - Gestió de la taula de fets: TFC_VEN.
 - Gestió del mòdul estadístic: TFC_ABA, TFC_ACA, TFC_ADMA, TFC_AHA, TFC_APA, TFC_INA.
- **Disparadors:** Totes les taules tenen disparador per enregistrar la informació d'usuari d'inserció, data i hora d'inserció, usuari de modificació i data i hora de modificació.
- **Seqüències:** Es crea la seqüència RAP_SEQ per mantenir la clau primària de la taula TFC_RASTRE_PROCES

En el punt "7.6 Annex 6. Documentació Tècnica" de la memòria s'adjunta la documentació tècnica dels objectes del Data Warehouse.

Mòdul estadístic.



Mòdul estadístic.

- **TFC_INDICADORS_ANY:** Conté els següents indicadors:
 - Indicadors per any. Cada registre està associat un any en concret.
Els indicadors per any són:
 - Benefici net total de la cadena per any.
 - Percentatge de benefici net total obtingut per tendes virtuals per cada any
 - Indicadors històrics. Cada registre associat un any en concret, es guarden els indicadors històrics fins l' any associat al registre.
Els indicadors històrics són:

- **TFC_AGR_BOT_ANY:** Conté els següents indicadors:
 - Indicadors per any: Es guarda la suma de benefici net per botiga i any. S' identifica si es tracta de la botiga amb més o menys vendes.
 - Indicadors històrics: Es guarda la suma de benefici net per botiga i els anys menors o iguals al associat al registre. S' identifica si es tracta de la botiga amb més o menys vendes de forma històrica.

Mòdul estadístic.

- **TFC_AGR_DIAMES_ANY:** Conté els següents indicadors:
 - Indicadors per any: Es guarda la suma de la quantitat venuda per dia del mes i any. S' identifica si es tracta del dia del mes amb més o menys vendes.
 - Indicadors històrics: Es guarda la suma de la quantitat venuda per dia del mes i els anys menors o iguals al associat al registre. S' identifica si es tracta del dia del mes amb més o menys vendes de forma històrica.
- **TFC_AGR_HOR_ANY :** Conté els següents indicadors:
 - Indicadors per any: Es guarda la suma de la quantitat venuda per hora del dia i any. S' identifica si es tracta de la hora del dia amb més o menys vendes.
 - Indicadors històrics: Es guarda la suma de la quantitat venuda per hora del dia i els anys menors o iguals al associat al registre. S' identifica si es tracta de la hora del dia amb més o menys vendes de forma històrica.

Mòdul estadístic.

- **TFC_AGR_CIU_ANY:** Conté els següents indicadors:
 - Indicadors per any: Es guarda la suma de benefici net per ciutat i any. S' identifica si es tracta de la ciutat amb més o menys vendes.
 - Indicadors històrics: Es guarda la suma de benefici net per ciutat i els anys menors o iguals al associat al registre. S' identifica si es tracta de la ciutat amb més o menys vendes de forma històrica.
- **TFC_AGR_PRO_ANY:** Conté els següents indicadors:
 - Indicadors per any: Es guarda la suma de benefici net per producte i any. S' identifica si es tracta del producte amb més o menys vendes.
 - Indicadors històrics: Es guarda la suma de benefici net per producte i els anys menors o iguals al associat al registre. S' identifica si es tracta del producte amb més o menys vendes de forma històrica.

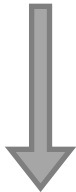
Càlcul del mòdul estadístic.

La solució que proposada i implementada, com es veurà en posteriors diapositives, es pot parametritzar. En funció de la parametrització, el càlcul del mòdul estadístic es podrà fer amb un dels següent mètodes:

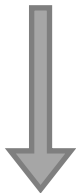
- **Càlcul manual dels indicadors del mòdul estadístic:** Es pot fer en qualsevol moment sota demanda cridant al següent procediment:
“TFC_INA.CALCULAR_INDICADORS_TOT”
El càlcul manual s’ haurà de fer obligatòriament si s’ ha configurat el paràmetre “CALC_IND_PER_CADA_VENDA” amb el valor “N”.
- **Càlcul automàtic incremental dels indicadors del mòdul estadístic:** Es detalla en presentacions posteriors.
- **Càlcul automàtic i total dels indicadors del mòdul estadístic:** Es detalla en presentacions posteriors.

- Paràmetre “**CALC_IND_PER_CADA_VENDA**” configurat amb el valor “**S**”.
- Paràmetre “**TIPUS_CALCUL_INDICADORS**” configurat amb el valor “**INC**”.

Alta / Modificació / Eliminació



TFC_VENDES



Taules del mòdul estadístic

Es generen o actualitza un únic registre de les taules TFC_AGR* del mòdul estadístic.

Es calcula la informació de la taula TFC_INDICADORS_ANY.

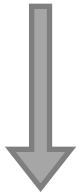
No es realitzen consultes complexes amb funcions d'agregats *sum* amb *group by*.

Els càlculs es fan únicament amb la informació del registre de venda.

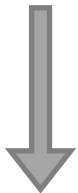
No es pot executar sota demanda.

- Paràmetre “**CALC_IND_PER_CADA_VENDA**” configurat amb el valor “**S**”.
- Paràmetre “**TIPUS_CALCUL_INDICADORS**” configurat amb el valor “**TOT**”.

Alta / Modificació / Eliminació



TFC_VENDES



Taules del mòdul estadístic

Per cada venda, s’ elimina la informació de l’ any i es torna a calcular a partir de la informació de la taula TFC_VENDES.

Es generen els registres de l’ any tractat a les taules TFC_AGR* del mòdul estadístic.

Es calcula la informació de la taula TFC_INDICADORS_ANY.

Es fan servir consultes amb funcions d’ agregats *sum* amb *group by* .

Els càlculs es fan únicament amb les vendes de l’ any associat al registre inserit a la taula TFC_VENDA.

Es pot executar sota demanda:

“TFC_INA.CALCULAR_INDICADORS_TOT”

Aspectes de valor afegit implementats.

En la solució desenvolupada s' han implantat els següents aspectes de valor afegit:

- **Particionament de la taula de fets:** Es faciliten els scripts necessaris per particionar la taula de fets.
- **Desenvolupament d' una aplicació en Apex:** S' ha implementat una aplicació en Apex. Aquesta aplicació es desenvolupa a mode d' exemple de com explotar les dades del DWH.
- **Càrrega múltiple de dades a partir de fitxers:** S' han implementat processos per permetre la càrrega múltiple d' informació a partir de fitxers. Es descriu en diapositives posteriors.
- **Solució parametritzada:** La solució té varis aspectes que es poden configurar (enviament de mails, modalitat de càlcul dels indicadors, ...)

Càrrega de dades múltiple.

Estructura de directoris.

└─ TFCBD	→	Ruta de instal·lació
└─ Documents	→	Documentació del projecte
└─ Instal·lacio	→	Directoris accessibles per la Base de Dades i scripts d'instal·lació
└─ Dades	→	Directoris accessibles per la Base de Dades
└─ Pendants_Processar	→	Fitxers pendents de processar
└─ ALT	→	Ruta on s'han de deixar les altes pendents de processar
└─ BAI	→	Ruta on s'han de deixar les baixes pendents de processar
└─ MOD	→	Ruta on s'han de deixar les modificacions pendents de processar
└─ Processats_ERR	→	Conté els fitxers amb els registres que han generat errors
└─ ALT	→	Ruta on el sistema deixa les altes que han donat errors
└─ BAI	→	Ruta on el sistema deixa les baixes que han donat errors
└─ MOD	→	Ruta on el sistema deixa les modificacions que han donat error
└─ Processats_OK	→	Conté els fitxers processats correctament
└─ ALT	→	Ruta on el sistema deixa els fitxers d' altes processats
└─ BAI	→	Ruta on el sistema deixa els fitxers de baixes processats
└─ MOD	→	Ruta on el sistema deixa els fitxers de modificacions processats
└─ scripts	→	Conté els scripts d'instal·lació dels components del DWH

Càrrega de dades múltiple.

Nom i format dels fitxers de dades.

Els camps dintre dels fitxers s' han de separar amb el caràcter '|'

ciutats.dat:

- Identificador de la ciutat.
- Nom de la ciutat.
- Identificador de la regió.

botigues.dat :

- Identificador de la botiga.
- Nom de la botiga.
- Identificador de la ciutat.
- Número de treballadors.
- E-Mail de contacte
- És franquícia?: S/N
- És virtual?: S/N
- Data de alta amb el format YYYYMMDD
- Data de baixa amb el format YYYYMMDD

regions.dat:

- Identificador de la regió.
- Nom de la regió.

productes.dat:

- Codi EAN 13 del producte.
- Nom del producte.
- Descripció del producte.
- Està actiu?: S/N.
- Data d' alta en el catàleg amb el format YYYYMMDD.
- Data de baixa del catàleg amb el format YYYYMMDD.

vendes.dat:

- Data i hora de la venda amb el format YYYYMMDDHH24MISS.
- Identificador de la botiga.
- Codi EAN 13 del producte.
- Quantitat venuda.
- Benefici net.
- Import brut.

Càrrega de dades múltiple.

Execució de la càrrega múltiple

1. Deixar els fitxer de dades a la ruta “Pendants_Processar”:

- Els fitxers han de complir el format adequat.
- Els fitxers s’ han de deixar en la ruta adequada en funció del tipus d’ operació (alta/modificació/baixa)

2. Executar l’ script de càrrega de dades:

```
DECLARE
```

```
  vRsp      varchar2(32000);  
  vIdeRap   number;
```

```
BEGIN
```

```
TFC_LDR.PROCESSAR_ALTES(vIdeRap, vRsp);  
dbms_output.put_line('Identificador de rastre per les altes: '||vIdeRap);  
dbms_output.put_line('Resultat de processar les altes: '||vRsp);
```

```
TFC_LDR.PROCESSAR_MODIFICACIONS(vIdeRap, vRsp);  
dbms_output.put_line('Identificador de rastre per les modificacions: '||vIdeRap);  
dbms_output.put_line('Resultat de processar les modificacions: '||vRsp);
```

```
TFC_LDR.PROCESSAR_BAIXES(vIdeRap, vRsp);  
dbms_output.put_line('Identificador de rastre per les baixes: '||vIdeRap);  
dbms_output.put_line('Resultat de processar les baixes: '||vRsp);
```

```
END;
```

3. Revisar rastre: TFC_RASTRE_PROCES, TFC_RASTRE_LINIA

Propostes de millora

Per a futures versions, es proposa incloure les següents millores:

- **Us de vistes materialitzades:** Permeten automatitzar alguns càlculs del mòdul estadístic.
- **Redundància de dades:** Revisar en quins casos pot ser útil tenir informació redundat per reduir el número de *joins* entre taules. El Data Warehouse passaria del model floc de neu a un model en estrella.
- **Us de *jobs*:** Programar certs càlculs en *jobs* permetrà reduir el temps de càrrega de la informació.
- **Nivells de traça de la informació:** Reduir la informació que es guarda en la taula de log. Permetre que es pugui parametritzar o només guardar els errors.
- **Enviament de mails:** Modificar els processos perquè enviïn mails en cas d'errors.